Visualisation of High-dimensional Data for Very Large Data Sets

David Wong[1]; Iain Strachan[2]; Lionel Tarassenko[1];

[1]Institute of Biomedical Engineering, University of Oxford
[2]Oxford Biosignals Ltd., Brook House, 174 Milton Park, Abingdon, Oxfordshire, UK

1. Introduction

Motivation

In the field of patient monitoring in critical care, researchers are often overwhelmed with large quantities of high-dimensional data. Initial exploration and analysis of such high-dimensional data is a difficult task. Any analytic tools must deal with the data in a coherent and intuitive manner on order to provide useful insight, but must also be usable with large volumes of data.

Sammon Maps

One popular method of visualising data is using a Sammon map. This produces a plot that attempts to keep the Euclidean distances between all pairs of data point in the 2D visualisation space as close as possible to those in the high-dimensional space. Mathematically, this is the same as minimising the Sammon STRESS objective function for N data samples:

$$STRESS = \frac{1}{\sum_{i=1}^{N}\sum_{j>i}^{N} d^*_{ij}} \sum_{i=1}^{N}\sum_{j>i}^{N} \frac{(d_{ij}-d^*_{ij})^2}{d^*_{ij}}$$

where the Euclidean distances between patterns *i* and *j* in the data space are denoted by d^*_{ij} and the corresponding points in visualisation space are denoted by d_{ij}

The Problem

One major problem with Sammon's algorithm is that creating Sammon maps is intractable for large data sets due to memory and computation time constraints as the STRESS calculation requires O(N^2) comparisons.

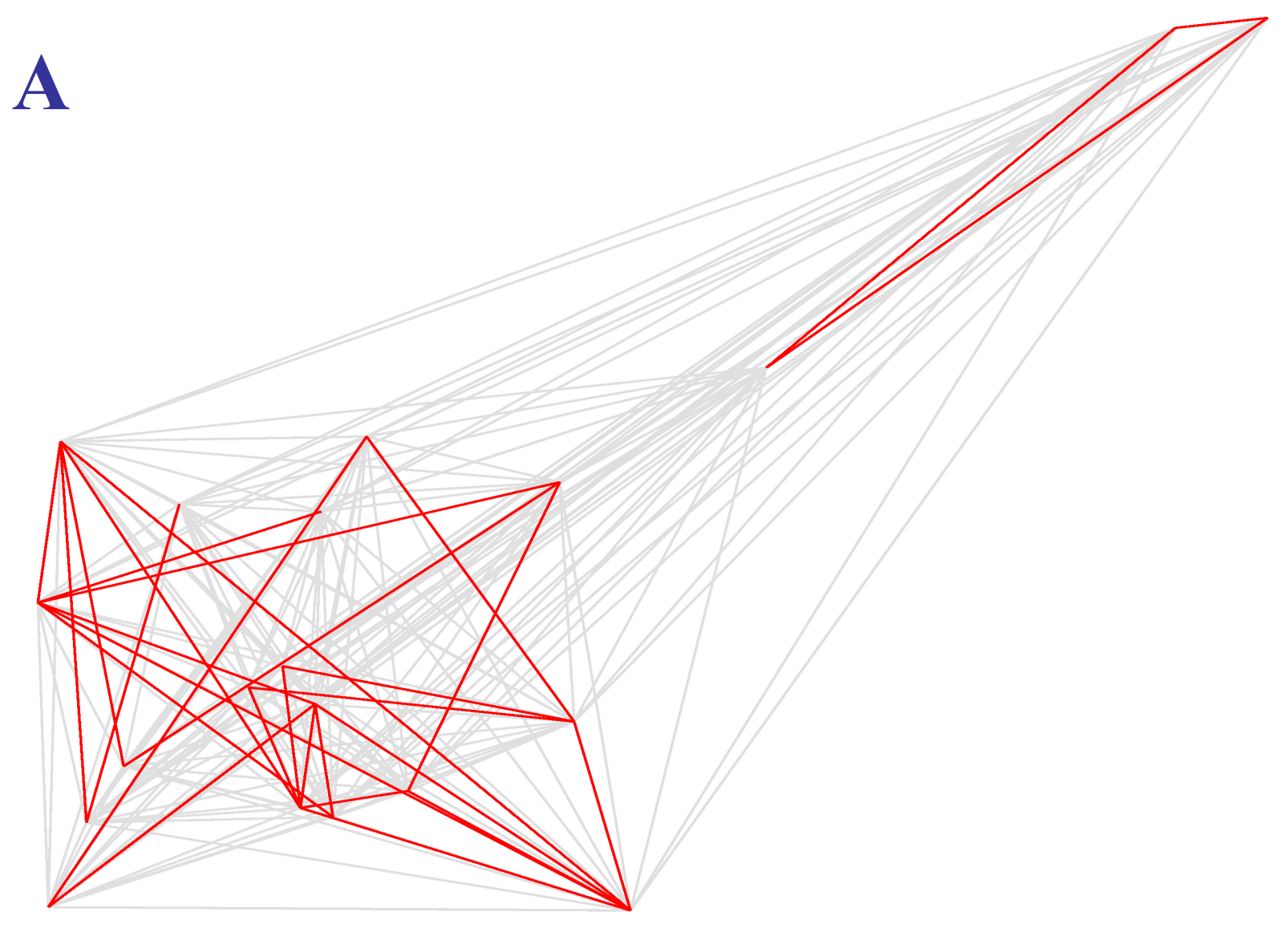
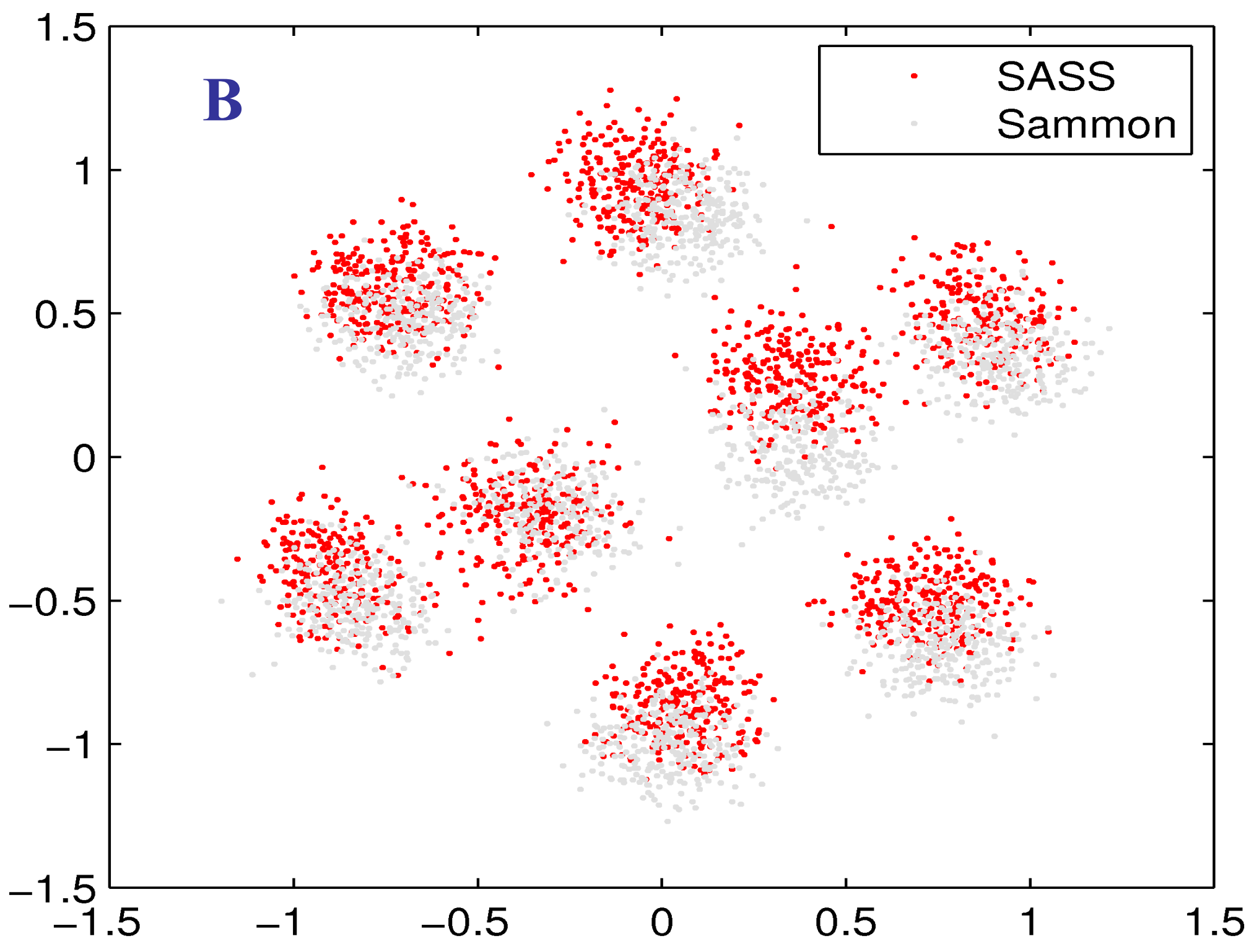


*Fig. 1(A-B). **A.)** shows the connectivity between data points for the original Sammon algorithm (grey) and for SASS (red). Each node represents a data point, and each edge represents an inter-point distance. In this example, two subsets have formed and SASS will fail to visualise the data correctly. Correct algorithm initialisation prevents this from occurring in practise. **B.)** shows a Sammon map generated using the STRESS and SASS metrics from a synthetic data set containing 20,000 points normally distributed at the corners of a 3D cube.*



2. Method

The Solution - SASS

We propose a novel alternative to the original Sammon Map algorithm which we have named the Sparse Approximated Sammon STRESS (SASS). SASS reduces the mapping objective function to one of order O(N) by sampling from the complete set of inter-point distance pairs. The modified STRESS objective function is:

$$SASS = \frac{1}{\sum_{i,j\in S} d^*_{ij}} \sum_{i,j\in S} \frac{(d_{ij}-d^*_{ij})^2}{d^*_{ij}}$$

SASS is then minimised by adjusting points in visualisation space. A pictorial comparison between SASS and STRESS calculations is shown in the Figure 1A, and an example output from both is shown for a synthetic data set in Figure 1B. For this medium sized data set, SASS performs as well as the original algorithm.

Initialisation of d_{ij} in Visualisation Space

Points in visualisation space are initialised using a two-stage approach. Firstly, a subset of the data is mapped using SASS. Secondly, the remaining data is approximately mapped to visualisation space using a linear transform to provide estimates of d_{ij}

Initialisation of Subset S

The set of inter-point distance pairs, S, was chosen to provide an equal proportion of 'local' and 'distant' connections to maintain good local and global structure in the mapping. To do this, data is clustered using K-means. A connection is 'local' if two vectors are from the same cluster, and 'distant' if vectors are selected from different clusters.

3. Results

SASS for Vital Sign Visualisation

Figure 3 was generated using data from a clinical trial at the University of Pittsburgh Medical centre (UPMC), which contained vital sign recordings taken over an eight-week period for a total of 300 patients. Four vital signs, the heart rate, breathing rate, arterial-oxygen saturation and blood pressure, were recorded simultaneously. In total 961,031 4D vital sign vectors were recorded, which corresponds to 28,782 hours of data collection.

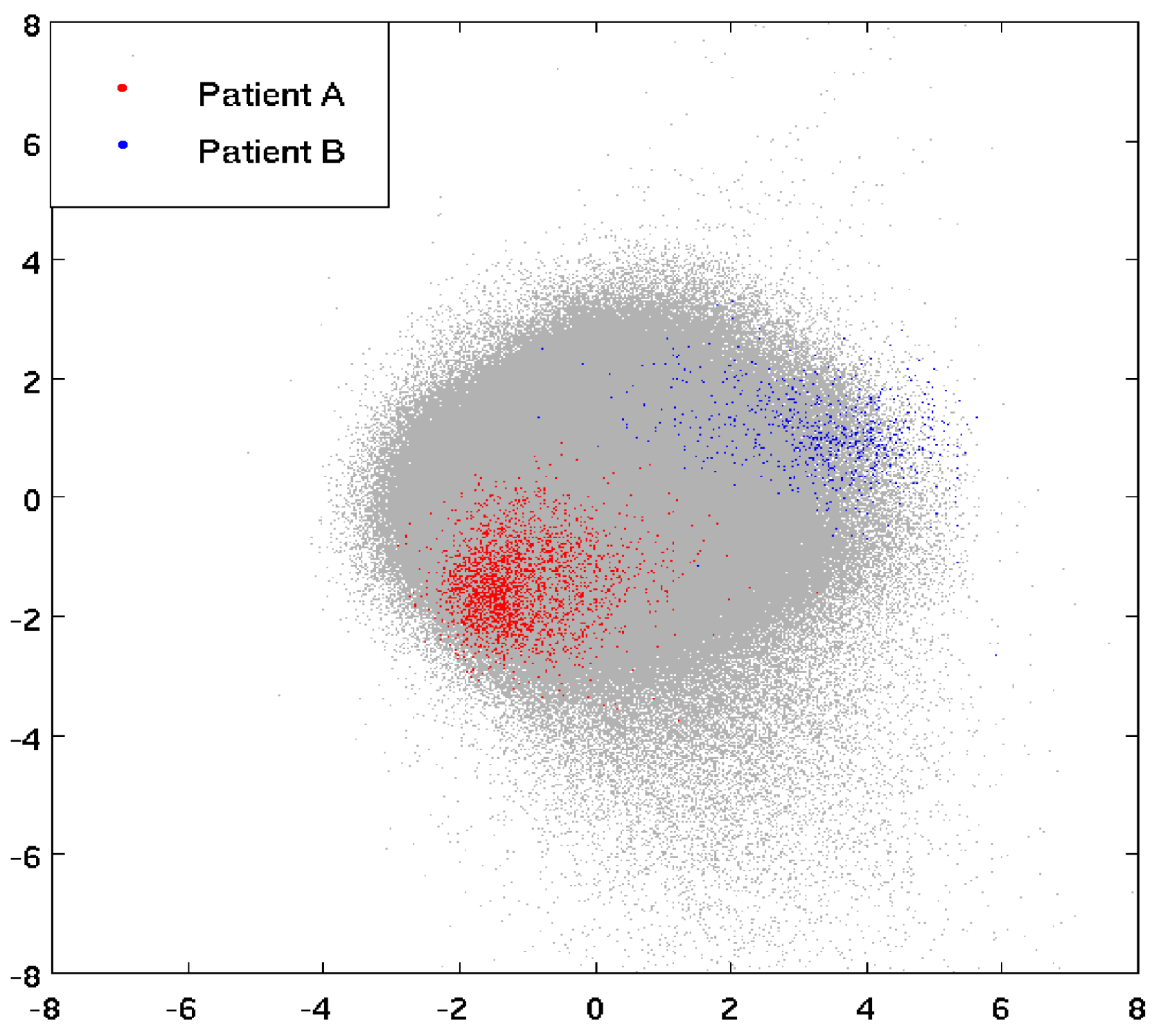


Fig. 3. A Sammon Map for the UPMC vital sign data. The whole data set, consisting of 961,031 4D vectors, is visualised in grey. Points corresponding to individual patients A and B are plotted in red and blue respectively.

For both patients, the visualised vital sign clusters are much more compact than the global model.

Figure 3 shows us that there may can be considerable variation in vital signs may vary from patient-to-patient. External factors such as patient age may cause this, and so, under certain circumstances, patient-specific models of vital sign data may be more appropriate than a global model of normality.

SASS for Drug Analysis

Figure 4 was constructed from a data set of 8867 ECG waveforms from the first eight hours of a 24 hour recording during a clinical study of the drug D-sotalol. Each point on the plot represents wavelet coefficients from the ECG waveforms, and the 'inter-point distance' is a similarity measure that is calculated using Dynamic Time Warping.

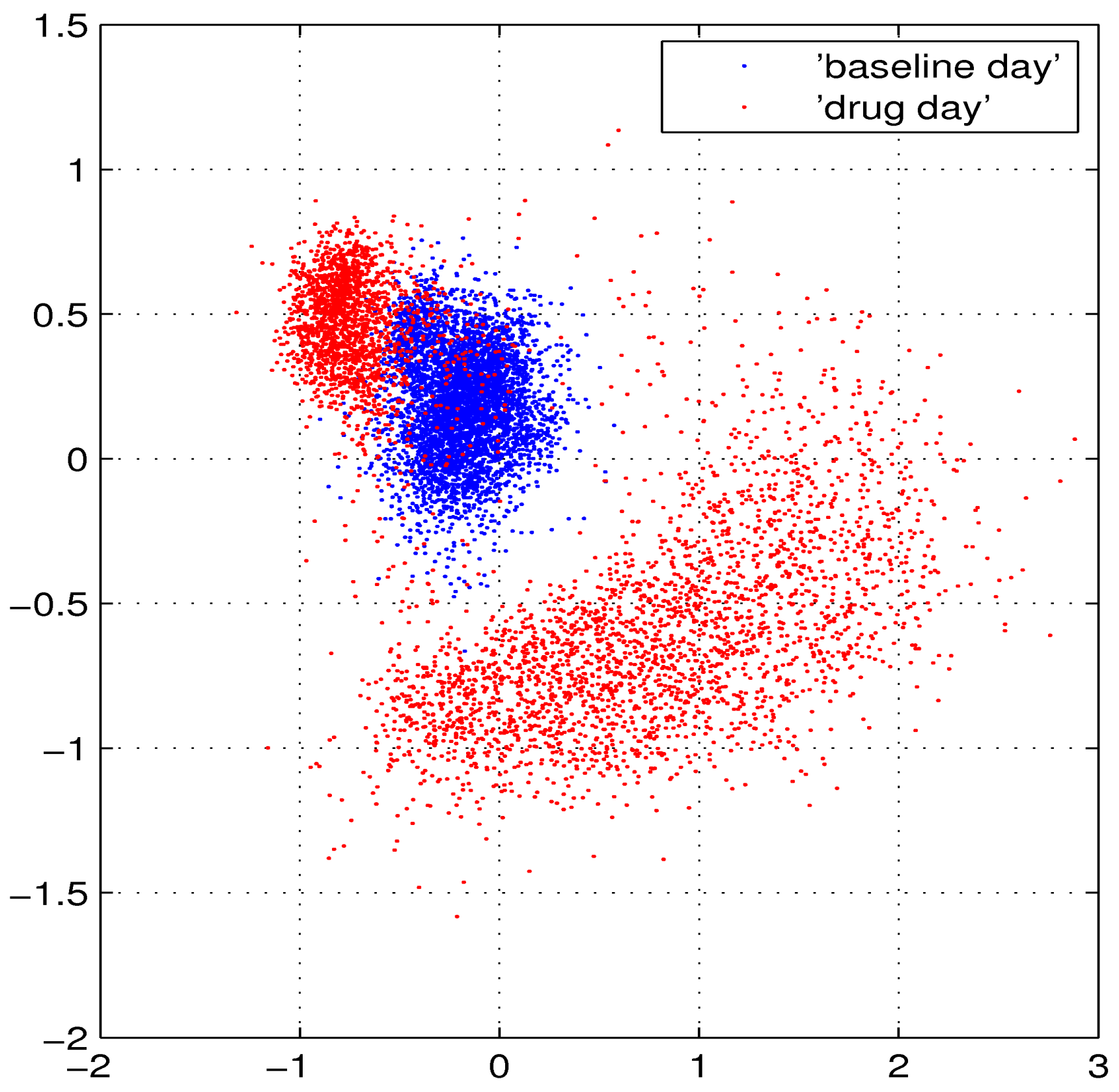


Fig. 4. A Sammon Map showing the effect of the drug D-sotalol on Electrocardiogram waveform morphology. The `baseline' day where no drug was taken is mapped in blue, and the red points represent recordings following the administration of the drug.

The more compact red clusted corresponds to readings taken immediatletly after drug delivery, while the highly scattered cluster represents points after a further 2 hours.

It is known that D-sotalol produces large changes in the ECG waveform morphology, especially in the region of Ventricular repolarisation (T-wave). The visualisation clearly shows a big effect from the drug, confirming the morphology change effect.